Постановка задачи

- По данным из файла проверить, является ли ряд стационарным в широком смысле.
- Провести визуальную оценку, отрисовав ряд и скользящую статистику.
- Провести тест Дики – Фуллера.
- Оценить достоверность статистики.
- Разложить временной ряд на тренд, сезональность, остаток в соответствии с аддитивной, мультипликативной моделями.
- Проверить, является ли ряд интегрированным порядка k. Если да, то применить модель $ARIMA$.
- Визуализировать решение

Подход к решению

Теоретический материал

Определение. *Выборка* – это часть объектов из генеральной совокупности, отобранных для изучения, с целью получения информации обо всей генеральной совокупности.

Определение. Пусть задано вероятностное пространство $(\boldsymbol{\Omega}, \boldsymbol{\mathcal{F}}, P)$.
Случайной величиной называется $\xi : \boldsymbol{\Omega} \to \mathbb{R}$, т. ч. ξ – измерима, то есть

$$\xi^{-1}(B) = \{\, \omega \in \boldsymbol{\Omega} : \xi(\omega) \in \mathrm{B} \,\} \in \boldsymbol{\mathcal{F}}, \qquad \forall\ \mathrm{B} \in \mathcal{B}$$

Определение. *Временным рядом* называется совокупность значений какого-либо показателя за несколько последовательных периодов времени.

Параметрами временного ряда являются :

1. Период времени
2. Уровни ряда – значения показателя

Мы будем рассматривать временной ряд как выборку из последовательности случайных величин X_t , где t принимает целочисленные значения от 1 до T.

Определение. Совокупность случайных величин $\{\, X_t,\ t \in [1, T] \,\}$ будем называть *дискретным случайным или стохастическим процессом*.

Иногда говорят, что стохастический процесс «для каждого случая» является некоторой функцией времени, что позволяет рассматривать процесс как случайную функцию времени $X(t)$. При каждом фиксированном t значение стохастического процесса рассматривается просто как случайная величина. Эти два эквивалентных подхода позволяют рассматривать стохастический процесс как функцию двух величин, случая и момента времени: $X(\omega, t)$. При фиксированном случае у нас есть некоторая

последовательность значений первой случайной величины, второй, третьей и так далее, которую мы будем называть реализацией случайного процесса.

Определение. *Функцией распределения случайного вектора* $\bar{\xi} = (\xi_1, \xi_2, \dots, \xi_n)$ или *совместным распределением* случайных величин $\xi_1, \xi_2, \dots, \xi_n$ называется функция, определённая равенством

$F_{\bar{\xi}}(\bar{x}) = F_{\xi_1,\xi_2,\dots,\xi_n}(x_1, x_2, \dots, x_n) = P(\xi_1 < x_1, \xi_2 < x_2, \dots, \xi_n < x_n)$, где $\bar{x} = (x_1, x_2, \dots, x_n)$.

Определение. Случайный вектор $\bar{\xi} = (\xi_1, \xi_2)$ называется *непрерывным случайным вектором*, если существует такая неотрицательная функция $p_{\bar{\xi}}(x_1, x_2)$, что для любого прямоугольника $\boldsymbol{\Omega}$ на плоскости $(\boldsymbol{x_1}, \boldsymbol{x_2})$ вероятность события $\bar{\xi} \in \boldsymbol{\Omega}$ равна

$P(\bar{\xi} \in \boldsymbol{\Omega}) = \iint_{\boldsymbol{\Omega}} p_{\bar{\xi}}(x_1, x_2)\, \mathrm{d}x_1 dx_2$, функция $p_{\bar{\xi}}(x_1, x_2)$ – *совместная плотность распределения.*

Поскольку случайный дискретный процесс $X(\omega, t)$ представляет собой совокупность случайных величин, то его характеристикой будет совместная функция распределения или функция плотности распределения (если плотность существует). При рассмотрении временного ряда число случайных величин велико и может быть бесконечным. Поэтому для задания случайного процесса понадобится совокупность функций распределения: $f_1(x_{t_1})$; $f_2(x_{t_1}, x_{t_2})$; $f_3(x_{t_1}, x_{t_2}, x_{t_3})$; … , где индексы у величин x_{t_1} и x_{t_2} означают, что одна случайная величина рассматривается в момент t_1, вторая – в момент t_2 и так далее, и у них есть совместная функция распределения. Совокупность функций $f_1, f_2, \dots$ согласована следующим образом: каждую функцию распределения размерности n можно получить из функции распределения размерности $n+1$, для этого надо проинтегрировать функцию большей размерности по всем значениям одной из переменных.

Определение. Случайный процесс называется *строго стационарным* (стационарным в узком смысле), если сдвиг во времени не меняет ни одну из функций плотности распределения. Это значит, что если ко всем моментам времени прибавить некоторую (целочисленную) величину, то сама функция плотности не изменится,

$f_n(x_{t_1}, \dots, x_{t_n}) = f_n(x_{t_1+\Delta}, \dots, x_{t_n+\Delta})$ для всех n, моментов времени $t_1, t_2, \dots, t_n$ и целочисленных Δ.

Рассмотрим математическое ожидание непрерывной случайной величины X_t:

$$E\{X_t\} = \int_{-\infty}^{+\infty} z\, f_1(z) dz = \mu\ ,$$

предполагаем, что этот интеграл сходится.

Если процесс стационарный, то для любого t подынтегральное выражение не меняется, а значит математическое ожидание не зависит от времени. Аналогичный результат справедлив для дисперсии стационарного процесса : $Var(X_t) = \sigma^2$.

Поскольку значения временного ряда в различные моменты времени зависимы между собой, то можно рассмотреть их ковариацию :

$$Cov(X_{t_1}, X_{t_2}) = \iint (x_{t_1} - \mu)(x_{t_2} - \mu) f_2(x_{t_1}, x_{t_2}) dx_{t_1} dx_{t_2},$$

которая в силу стационарности процесса будет зависеть лишь от одной переменной : разности $(t_1 - t_2)$.

Определение. Совокупность значений ковариаций при всевозможных значениях расстояния между моментами времени называется *автоковариационной функцией* случайного процесса.

Обозначив $Cov(X_{t_1}, X_{t_2}) = \gamma$, получим $\gamma(0) = Cov(X_{t_1}, X_{t_1}) = \sigma^2$.

Будем рассматривать функцию $\gamma(\tau)$ как всевозможные значения автоковариаций, где τ пробегает целочисленные значения от $-\infty$ до $+\infty$. Тогда коэффициент корреляции (то есть ковариация, делённая на корень из произведения двух дисперсий) будет иметь вид $\rho(\tau) = \frac{\gamma(\tau)}{\gamma(0)}$. Это выражение определяет *автокорреляционную функцию*.

Определение. Случайный процесс называется *слабо стационарным* (стационарным в широком смысле), если его математическое ожидание и дисперсия существуют и не зависят от времени, а автокорреляционная (автоковариационная) функция зависит только от разности значений $(t_1 - t_2)$.

Определение. *Скользящее среднее* (англ. moving average, MA) – семейство функций, значения которых в каждой точке равны среднему значению исходной функции за предыдущий период.

Определение. *Стандартное отклонение* – это показатель рассеивания значений случайной величины относительно её математического ожидания. Оно демонстрирует, на сколько в среднем отклонился ряд от своей средней вариации (то есть от среднего арифметического).

Определение. Скользящее среднее и стандартное отклонение называются *скользящими статистиками*.

Тест Дики – Фуллера

Определение. *Тестом Дики – Фуллера* называется методика, которая используется для анализа временных рядов для проверки на стационарность.

Определение. *Авторегрессионная модель* (AR-) модель - модель временных рядов, в которой значения временного ряда в данный момент линейно зависят от предыдущих значений этого же ряда. *Авторегрессионный процесс порядка p* (AR(*p*) – процесс) определяется следующим образом

$X_t = c + \sum_{i=1}^{p} a_i X_{t-i} + \varepsilon_t$, где $a_1, \dots, a_p$ – параметры модели (коэффициенты авторегрессии), c - константа (часто для упрощения берётся равной нулю), ε_t - белый шум.

Стационарность авторегрессионного процесса зависит от корней характеристического полинома $a(z) = 1 - \sum_{i=1}^{n} a_i z_i$, корни которого в общем случае являются комплексными числами. Для того, чтобы процесс был стационарным, достаточно потребовать $|z| > 1$.

Простейшим примером является авторегрессионный процесс первого порядка AR(1):

$$X_t = c + rX_{t-1} + \varepsilon_t ,$$

у которого коэффициент авторегрессии совпадает с коэффициентом автокорреляции. Тогда для него $a(z) = 1 - rz$, значит $z = \frac{1}{r}$ и условие стационарности примет вид $|r| < 1$.

Определение. Если характеристическое уравнение $a(z) = 0$ имеет корни, равные по модулю единице, то эти корни называются ***единичными***.

Таким образом, временной ряд имеет ***единичный корень***, если его первые разности образуют стационарный ряд. Обозначим $y(t) \sim I(1)$, если ряд первых разностей $\Delta y_t = y_t - y_{t-1}$ является стационарным, то есть $\Delta y_t \sim I(0)$, где $I(j)$ означает, что ряд является интегрированным порядка j, а $I(0)$ - что ряд стационарен.

При помощи теста Дики – Фуллера проверяют значение коэффициента a для AR(1) :

$y_t = ay_{t-1} + \varepsilon_t$, где y_t – временной ряд, а ε_t – ошибка. Если :

- $a = 1$, то процесс имеет единичный корень, а значит ряд y_t не стационарен ;
- $|a| < 1$, то ряд стационарный ;
- $|a| > 1$ не свойственно для временных рядов, которые встречаются в реальной жизни – требуется более сложный анализ.

Преобразуем уравнение $y_t = ay_{t-1} + \varepsilon_t$ в $y_t - y_{t-1} = ay_{t-1} - y_{t-1} + \varepsilon_t$, то есть

$\Delta y_t = (a-1)y_{t-1} + \varepsilon_t$, обозначим $b = a - 1$, тогда $\Delta y_t = by_{t-1} + \varepsilon_t$.

Тогда проверка гипотезы о единичном корне означает проверку нулевой гипотезы о равенстве нулю коэффициента b. Напомним, что ***нулевой (или основной) гипотезой*** является та, которая принимается верной, пока не доказано обратное, а ***альтернативной гипотезой***, - та, которая принимается в случае отклонения нулевой.

В нашей задаче :

- основная гипотеза $H_0: b = 0$ – процесс нестационарен;
- альтернативная $H_1: b < 0$ – процесс стационарен.

Как происходит проверка на наличие единичных корней?

Для получения ответа сравниваем найденное с помощью функции $\boldsymbol{sm.tsa.adfuller}(series)$ $p-value$ – достоверностью статистики с $critical\_values$ – критическими значениями с уровнем значимости 5% :

- $p-value > critical\_values,$ тогда единичный корень есть и ряд нестационарен;
- иначе ряд стационарен.

Определение. *Достоверность статистики* $(\boldsymbol{p-value})$ – мера уверенности в «истинности» результата. Чем меньше $\boldsymbol{p-value}$, тем больше доверия. Как только $\boldsymbol{p-value}$ становится больше определённого параметра – $\boldsymbol{critical\_values}$, то понимаем, что доверять нельзя и делаем вывод, что единичные корни есть.

Определение. *Уровень значимости* – показывает (обычно в процентном соотношении) степень отклонения от гипотезы. То есть, если наша достоверность превысила, например 5% уровень значимости, то гипотеза отвергается и значит процесс нестационарен.

Тренд, сезональность, остаток. Аддитивная и мультипликативная модели

Определение. *Трендом* $(\boldsymbol{T})$ временного ряда называется изменение его показателей (то есть общее направление графика ряда : возрастает, убывает или не изменяется).

Определение. *Сезональностью* $(\boldsymbol{S})$ временного ряда называются периодические колебания, наблюдаемые во временных рядах.

Определение. *Остатком* $(\boldsymbol{E})$ временного ряда называется разница между предсказанным и наблюдаемым значением.

Определение. Модели, в которых временной ряд представлен в виде суммы перечисленных компонент, называются *аддитивными*, в виде произведения – *мультипликативными*.

Аддитивная модель имеет вид : $\boldsymbol{Y = T + S + E}$, где $\boldsymbol{Y}$ – временной ряд.

Мультипликативная модель имеет вид : $\boldsymbol{Y = T * S * E}$.

Рассмотрим аддитивную модель.

Сезональность. Необходимо найти скользящее среднее, от него ещё раз найти скользящее среднее – получим центрированное скользящее среднее $(\boldsymbol{\overline{MA}})$. Тогда

$$\boldsymbol{S = Y - \overline{MA}}\ .$$

Тренд. Ищем тренд с помощью метода наименьших квадратов.

Определение. *Метод наименьших квадратов* $(\boldsymbol{МНК})$ – основан на минимизации суммы квадратов отклонений некоторых функций от искомых переменных.

Линейное уравнение тренда имеет вид $\boldsymbol{Y = a_0 t + a_1}$, тогда система уравнений МНК выглядит следующим образом :

$$\begin{cases} \boldsymbol{a_0 n + a_1 \sum t = \sum Y} \\ \boldsymbol{a_0 \sum t + a_1 \sum t^2 = \sum Y \cdot t} \end{cases},$$

где $\boldsymbol{n}$ – число промежутков времени.

Остаток. Находится по формуле $\boldsymbol{E = Y - T - S}$.

Теперь рассмотрим мультипликативную модель.

Сезональность. $\boldsymbol{S = \frac{Y}{MA}}$.

Тренд. Находится аналогично с аддитивной моделью с помощью МНК.

Остаток. $\boldsymbol{E = \frac{Y}{T \cdot S}}$.

Интегрированность

Определение. *Интегрированный временной ряд* – это нестационарный временной ряд, разности некоторого порядка от которого являются стационарным рядом.

Определение. Временной ряд X_t называется *интегрированным порядка* $\boldsymbol{k}$ (обозначается $X_t \sim I(k)$), если разности ряда $\boldsymbol{k}$ – го порядка $\Delta^{\boldsymbol{k}} \boldsymbol{y_t}$ – являются стационарными, в то время как разности меньшего порядка (включая нулевой порядок, то есть сам временной ряд) не являются стационарными рядами. В частности, $I(0)$ – стационарный процесс.

$$\Delta \boldsymbol{y_t = y_t - y_{t-1}}$$

$$\Delta^{\boldsymbol{2}} \boldsymbol{y_t = \Delta y_t - \Delta y_{t-1} = y_t - 2 y_{t-1} + y_{t-2}}$$

$$\Delta^{\boldsymbol{k}} \boldsymbol{y_t = \Delta^{k-1} y_t - \Delta^{k-1} y_{t-1}}$$

Модель $ARIMA$

Определение. *Модель скользящего среднего* (moving average model, MA) – в ней моделируемый уровень временного ряда можно представить как линейную функцию прошлых ошибок, то есть разностей между фактическими и теоретическими уровнями.

Модель скользящего среднего q –го порядка $MA(q)$ имеет вид : $X_t = \sum_{j=0}^{q} b_j \varepsilon_{t-j}$,

где ε_t – белый шум, b_j – параметры модели.

Определение. $ARMA$ (англ. autoregressive moving-average model) – математическая модель, используемая для анализа и прогнозирования стационарных временных рядов.

Обобщает две более простые модели: авторегрессии (AR) и скользящего среднего (MA). Моделью $ARMA(p,q)$, где p,q – целые числа, задающие порядок модели, называется следующий процесс генерации временного ряда $\{X_t\}$:

$$X_t = c + \varepsilon_t + \sum_{i=1}^{p} \alpha_i X_{t-i} + \sum_{i=1}^{q} \beta_i \varepsilon_{t-i} ,$$

где c – константа, $\{\varepsilon_t\}$ – ***белый шум***, то есть последовательность независимых и одинаково распределённых случайных величин с нулевым математическим ожиданием, $\alpha_1, \dots, \alpha_p$ и $\beta_1, \dots, \beta_q$ – действительные числа, авторегрессионные коэффициенты и коэффициенты скользящего среднего, соответственно.

Определение. ***ARIMA*** (англ. autoregressive integrated moving average) – интегрированная модель авторегрессии – скользящего среднего – модель анализа временных рядов. Является расширением моделей $ARMA$ для нестационарных временных рядов, которые можно сделать стационарными взятием разностей некоторого порядка от исходного временного ряда.

Модель $ARIMA(p,d,q)$ для нестационарного временного ряда X_t имеет вид :

$$\Delta^{\boldsymbol{d}} X_t = c + \varepsilon_t + \sum_{i=1}^{p} a_i \Delta^{\boldsymbol{d}} X_{t-i} + \sum_{j=1}^{q} b_j \varepsilon_{t-j} ,$$

где ε_t – стационарный временной ряд, c, a_i, b_j – параметры модели, $\Delta^{\boldsymbol{d}}$ – разностный оператор.

Таким образом, необходимо построить $ARIMA(p,d,q)$, где p – порядок $AR(p)$, q – порядок $MA(q)$, d – порядок интегрированности :

$$ARIMA(p,d,q) = AR(p) + MA(q) \sim I(d) .$$

Определение. Для нахождения порядка авторегрессионной модели временного ряда X_t используется *функция частичной автокорреляции* $pacf(k)$, которая имеет вид :

$$\boldsymbol{pacf(k)} = \begin{cases} \boldsymbol{corr(x_{t+k}, x_t)} , \boldsymbol{k = 1} \\ \boldsymbol{corr\left(x_{t+k} - x_{t+k}^{k-1}, x_t - x_t^{k-1}\right)}, \boldsymbol{k > 1} \end{cases},$$

где $\boldsymbol{x_t^{k-1} = \beta_1 x_{t+1} + \beta_2 x_{t+2} \dots + \beta_{k-1} x_{t+k-1}}$, $\boldsymbol{x_{t+k}^{k-1} = \beta_1 x_{t+k-1} + \beta_2 x_{t+k-2} \dots + \beta_{k-1} x_{t+1}}$.

Коэффициент детерминации

Определение. *Коэффициент детерминации* (R^2) – это доля дисперсии зависимой переменной, объясняемая рассматриваемой моделью зависимости. Другими словами, это квадрат коэффициента корреляции выборки.

Информационный критерий Акаике

Определение. *Информационный критерий* – применяемая в статистике мера относительного качества статистических моделей, учитывающая степень «подгонки» модели под данные с корректировкой на используемое количество оцениваемых параметров. Информационные критерии используются исключительно для сравнения моделей между собой, без содержательной интерпретации значений этих критериев. Обычно чем меньше значения критериев, тем выше относительное качество модели.

Определение. *Информационный критерий Акаике* (AIC) – критерий, применяющийся для только для выбора из нескольких статистических моделей.

$$\text{Общий вид}: \quad AIC = \frac{2k}{n} + ln\frac{S_{res}}{n} \ ,$$

где k – количество параметров модели, n – число наблюдений, S_{res} – остаток.

Реализация

Используемые библиотеки: $pandas, matplotlib, statsmodels, sklearn, seaborn$.

Чтение

Читаем данные из файлов $training.xlsx$ и $testing.xlsx$ с помощью функции $read\_excel$, указывая, что столбец $Date$ будет являться индексом для наших данных. Также указываем с помощью $asfreq('MS')$, что нашим периодом индекса будет начало месяца $(MS = month\ start)$.

Анализ стационарности

Чтобы понять, является ли наш ряд стационарным, будем использовать визуализацию скользящих статистик и тест Дики-Фуллера. Для визуализации скользящего среднего и скользящего стандартного отклонения разбиваем ряд на части функцией $rolling()$ и применим к ним, соответсвенно, функции $mean()$ и $std()$. Далее строим графики с помощью библиотеки $matplotlib$. Для проведения теста Дики-Фуллера используем функцию $adfuller()$, выбрав уровень значимости 0.05. Для данного ряда получаем, что значение $p - value$ намного больше 0.05 и, следовательно, мы не можем отклонить гипотезу о нестационарности ряда.

Декомпозиция

Раскладываем ряд на тренд, сезонность и остаток с помощью функции $seasonal\_decompose()$, изменяя параметр $model$ для аддитивной и мультипликативной моделей. Строим графики, применяя функцию $add\_subplot()$, чтобы сгруппировать их вместе. Анализируем остаток ряда, предварительно избавившись от пропущенных значений с помощью функции $dropna()$. Так как $p - value$ очень мало и статистика теста меньше, чем 5% уровень значимости, то мы можем отклонить нулевую гипотезу. Получаем, что остаток ряда - стационарный.

Построение модели

Дифференцируем ряд, вычитая из него значения, смещенные на один месяц функцией $shift()$. После первого дифференцирования ряд становится стационарным, что видно из графиков и теста, следовательно, его порядок интегрированности равен 1. Строим автокорреляционную и частичную автокорреляционную функции с помощью функций $plot\_acf()$ и $plot\_pacf()$ соответственно. По ним определяем параметры для модели $ARIMA(p, d, q)$. Параметр d равен порядку интегрированности, то есть 1. Параметр p - это номер последнего лага на графике частичной автокорреляции, значение на котором сильно отлично от нуля. Параметр q - количество сильно отличных от нуля значений на графике автокорреляции. Далее строим две различных модели и находим их коэффициенты с помощью функции $fit()$.

Предсказание

С помощью функции $predict()$ строим прдесказание для значений следующих месяцев и сравниваем их со значениями из файла $testing.xlsx$, вычисляя $r2\_score$.

Выводы

Заданный нам временной ряд нестационарен, так как стандартное отклонение и скользящее среднее зависят от времени. Тест Дики – Фуллера это подтверждает: достоверность статистики больше уровня значимости (больше $0{,}05$).

Для обеих моделей оригинальный ряд и тренд нестационарны, так как зависят от времени.

Сезонная составляющая и остаток (для обеих моделей) стационарны, так как их значения колеблются в малых окрестностях констант: для аддитивной модели – возле нуля; для мультипликативной – возле единицы (причём в этой модели окрестность колебания на порядок меньше).

Лучший результат, который может дать $R2\ score$ - это 1.0. Если результаты больше или меньше 1.0 то, это говорит об очень большой неточности предсказания модели. Поскольку у наших моделей значения 10.1235 и 10.3651 что говорит о том, что они не точные. Считается, что наилучшей будет модель с наименьшим значением критерия AIC, и в нашем случае это модель $ARIMA(1, 1, 4)$ со значением 246.

Работу выполнили

Семёнов Андрей, Арбузов Пётр, Кюнченкова Дарья, 312 группа.

Вклады участников в работу

Семёнов Андрей, Арбузов Пётр: написание кода.
Кюнченкова Дарья: написание readme.